



**Federal Aviation
Administration**

DOT/FAA/AM-05/23
Office of Aerospace Medicine
Washington, DC 20591

Comparison of a Typical Electronic Attitude-Direction Indicator With Terrain-Depicting Primary Flight Displays for Performing Recoveries From Unknown Attitudes: Using Difference and Equivalence Tests

Dennis B. Beringer
Jerry D. Ball
FAA Civil Aerospace Medical Institute
Oklahoma City, OK 73125
Kelly Brennan
Sitafa Taite
School of Industrial Engineering
University of Oklahoma
Norman, OK 73019

December 2005

Final Report

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents thereof.

Technical Report Documentation Page

1. Report No. DOT/FAA/AM-05/23		2. Government Accession No.		3. Recipient's Catalog No.	
4. Title and Subtitle Comparison of a Typical Electronic Attitude-Direction Indicator With Terrain-Depicting Primary Flight Displays for Performing Recoveries From Unknown Attitudes: Using Difference and Equivalence Tests				5. Report Date December 2005	
				6. Performing Organization Code	
7. Author(s) Beringer DB ¹ , Ball JD ¹ , Brennan K ² , Taite S ²				8. Performing Organization Report No.	
9. Performing Organization Name and Address ¹ FAA Civil Aerospace Medical Institute P.O. Box 25082 Oklahoma City, OK 73125 ² School of Industrial Engineering University of Oklahoma Norman, OK 73019				10. Work Unit No. (TRAIS)	
				11. Contract or Grant No.	
12. Sponsoring Agency name and Address Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591				13. Type of Report and Period Covered	
				14. Sponsoring Agency Code	
15. Supplemental Notes Work was accomplished under approved task AHRR521					
16. Abstract A study was conducted to determine if primary flight displays (PFDs) depicting terrain could be used with a level of safety equivalent to electronic attitude-direction indicators (EADIs) without terrain. Five groups of 8 pilots each flew scenarios in a flight simulator using one of three PFDs (EADI, full-color terrain, uniformly brown terrain) with or without guidance cues. Performances of recoveries from unknown attitudes using the EADI were measured first as a baseline, followed by trials with one of the experimental formats. Performance measures included initial response time, total recovery time, and both initial and secondary control reversals. Traditional "difference" analyses found no significant performance differences between groups. Analyses using confidence intervals to assess equivalence of distributions showed that group performances were practically equivalent. Pilot preferences were examined and are reported. It was concluded that the specific terrain representations examined provided for performance at least equal to if not better than the conventional EADI. This comparative technique is recommended for situations in which one wishes to demonstrate that a proposed device or system is no worse than or roughly equivalent to something already in use.					
17. Key Words Primary Flight Display, Terrain Display, Unknown Attitudes, EFIS, Equivalence Tests				18. Distribution Statement Document is available to the public through the Defense Technical Information Center, Ft. Belvoir, VA 22060; and the National Technical Information Service, Springfield, VA 22161	
19. Security Classif. (of this report) Unclassified		20. Security Classif. (of this page) Unclassified		21. No. of Pages 11	
				22. Price	

ACKNOWLEDGMENTS

The authors thank Barry Runnels for his assistance with the operation of the AGARS during the study and Kevin Williams for his programming of the calculations and graphical depictions outlined in Rogers et al. (1993).

COMPARISON OF A TYPICAL ELECTRONIC ATTITUDE-DIRECTION INDICATOR WITH TERRAIN-DEPICTING PRIMARY FLIGHT DISPLAYS FOR PERFORMING RECOVERIES FROM UNKNOWN ATTITUDES: USING DIFFERENCE AND EQUIVALENCE TESTS

BACKGROUND

It is frequently necessary to determine whether a system submitted for certification provides for a level of safety equivalent to the system(s) that it is proposed to replace. In most cases, practicality limits the number of individuals who can participate in airborne testing of the system, and data collection is sometimes, of necessity, limited to the use of checklists with evaluation criteria. Even when quantitative objective laboratory evaluations in flight simulators are employed, the efforts often culminate in an attempt to demonstrate that something is *better* than something else, not that something is *equivalent* to something else. However, we often want to say that something *is* equivalent to something else, particularly when we are attempting to retain a standard level of system performance. It is not, obviously, sufficient to fail to reject the null hypothesis (no difference) and then claim to have proven it. On the other side of the coin, it is also not necessary to demonstrate that a proposed system produces significantly *better* performance than the standard. Rather, it should be sufficient to be unable to detect a difference while also making some statement about how closely performance approximates a standard (equivalence). An approach that could be used for “equivalence” testing involves confidence intervals (Rogers, Howard, & Vessey, 1993; Reising, Liggett, Kustra, & Snow, 1998; Seaman & Serlin, 1998), and its implementation is shown for the following specific example.

One major component of Electronic Flight Instrumentation Systems (EFIS) is the Primary Flight Display (PFD). Although PFDs initially depicted attitude and flight-guidance information, they now can include forward-looking perspective views of both guidance information (Beringer, 2000) and of the outside world (Alter, Barrows, Jennings, & Powell, 2000). Data relevant to GA are available that may be useful for determining what the allowable range of variation in design parameters can be that still supports effective pilot performance. A series of studies was performed at the NASA Langley Research Center examining various terrain representations, and assessing pilot preferences for field of view and style of depiction (Arthur, Prinzel, Kramer, Parrish, & Bailey, 2004). One issue not specifically addressed was recovery from unknown or unusual attitudes. This specific concern

was addressed in one certification process by requiring that the terrain depiction be removed from the PFD when the aircraft exceeded certain pitch or roll criteria because of concern that the depicted terrain might cause confusion or interfere with recovery. However, there were no empirical data to indicate what role, positive or negative, the terrain depiction might play in the recoveries.

A study was conducted to determine (1) if pilots would recover to the terrain horizon rather than the zero-pitch line if the two were different, as would be seen in mountainous terrain, (2) if positive guidance cues (Gershzhohn, 2001) could ameliorate any potential negative effects of the terrain background and (3) if the coloration of the terrain presentation would affect performance. A second step was added to the final analyses to determine if a statistically based comparison could be used here, as was by Reising et al. (1998), to assess equivalence of performances.

METHOD

Experimental Display Formats

The *baseline* display was a conventional Electronic Attitude-Direction Indicator (EADI; blue sky, brown ground) with airspeed, altitude and vertical speed shown in tape format along the left and right edges of the display with a compass card at the bottom of the display (Figure 1). Two formats were added that depicted a terrain background, one in full color and one with blue sky and uniformly brown ground (simulating a PFD that has already been certified). Additionally, guidance arrows were added to the baseline and full-color formats, producing five combinations of ADI format and guidance information. Pitch arrows were linear (Figure 2), appearing when the aircraft attitude was greater than 15 degrees up or 13 degrees down and disappearing when the aircraft was within 5 degrees of zero pitch, pointing from the aircraft symbol to the horizon. Roll arrows (Figure 1) were of arc form, appearing when the aircraft exceeded 25 degrees of bank and disappearing when the aircraft was within 10 degrees of zero bank, pointing from the plane of the wings to the horizon line. The roll-command arrow took precedence over the pitch-command arrow when the aircraft was pitched down, and the priority was reversed when the aircraft was pitched up.

Terrain was based on variable-sized polygons anchored to elevation posts with photo-realistic texture or uniform brown texture applied to the polygons. The following illustrations show the PFD in full-color terrain mode with pitch-guidance arrow (Figure 2) and in EADI mode with roll-guidance arrows (Figure 1).

Horizon line. The horizon line was constructed such that it would have high contrast against the vast majority of possible backgrounds. This is not normally an issue with traditional head-down attitude direction indicators (ADIs), as the horizon on these displays is represented as the boundary between differently colored filled areas, often with a line of a different color between them. It is also possible to use a single-color line (as long as it conforms to MIL-STD-1787C, 5.1.2.1; Horizon reference; the standard does not deal specifically with terrain-depicting PFDs, nor does the SAE Aerospace Recommended Practice document on perspective displays deal specifically with this horizon-line issue) in terrain-depicting displays where the ground and sky representations are of known uniform colors (i.e., the Chelton display uses a uniformly brown ground and

blue sky). Our horizon line consisted of three two-pixel bands alternating black-white-black on a 640 by 480 display field, consistent both with horizon lines used in other full-color terrain display experiments and with general guidelines. The PFD image was enlarged to 800 by 600 pixels in the simulator and was approximately 7.5 inches wide (16.4 degrees) by 5.6 inches tall (12.3 degrees) (a 9.38 inch diagonal). The display portrayed about 30 degrees laterally of the outside world.

Experimental Design

A three-factor mixed design was employed, with terrain background (full-color or none) and guidance arrows (present or absent) as the *between* factors and trial block as the *within* factor. A supplemental condition, brown-only terrain, was added after contribution of guidance arrows had been assessed (Figure 3). Dependent variables included initial response time (IRT; time to first control input), total recovery time (TRT), primary control-input reversals (first response in wrong direction), and secondary control-input reversals (subsequent response in wrong direction).



Figure 1. EADI with roll-recovery arrows shown.



Figure 2. Terrain-depicting PFD (full color) with pitch-recovery arrow.

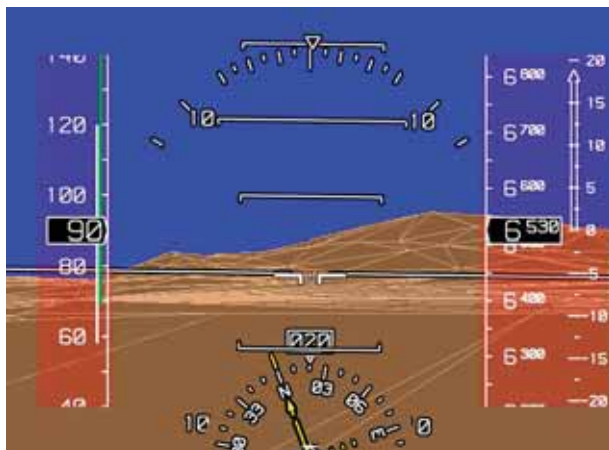


Figure 3. Supplemental brown-terrain condition.

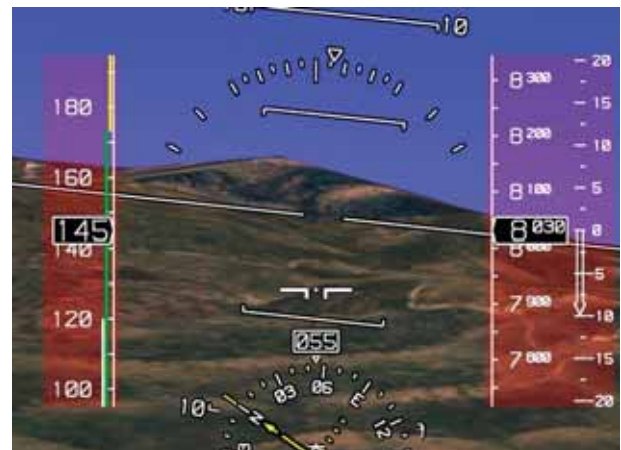


Figure 4. Full-color terrain PFD showing Sandia Peak in field of view.

Two *sampling variables, terrain depiction at roll-out and attitude at recovery onset*, were used to vary conditions; half of the trials ended facing the mountains and half facing a lower-elevation plain; attitudes used variation in pitch (+20, 0, and -15 degrees) and bank (60 degrees left, 0, 60 degrees right) (zero-zero excepted). Three supplemental trials were added for approximately the last 7 pilots in each group, including a near-mountains trial (peak extending to +5 degree pitch reference), an inverted trial (between 150- and 160-degree bank), and a 40-degree displayed field-of-view trial (to determine if pilot preference for a wider displayed field of view was associated with any performance advantage). A view showing Sandia Peak in the displayed field of view appears in Figure 4, but the peak is not depicted reaching the 5-degree pitch-up line as was the criterion for setting up a recovery trial.

Equipment and Participants

Data were collected using a fixed-base simulator configured to represent a Piper Malibu. Given that (1) the usual practice in training for unknown attitude recoveries is to “confuse” the participant by providing misleading vestibular cues and that (2) real-world unusual attitudes are often entered slowly, without awareness, and without overt acceleration cues until well into a maneuver, the absence of motion appeared an equally acceptable condition to impoverish pilot perception as to the state of the aircraft. The PFD was presented on a LCD on the left side of the instrument panel directly in front of the participant, and the experimenter-pilot (EP) flew from the right seat with a repeater display. The out-the-window view at altitude represented a hard-IFR situation with no visible environmental cues (gray). Performance data were recorded digitally, with supplemental audio and video data recorded to DVD.

Participants were 40 GA pilots (38 male, 2 female) recruited from the local community, 8 assigned to each of the 5 display conditions. Age and overall flight hours were balanced across groups as participants entered the experiment. Ages ranged from 19 to 57 years. All were certified minimally as Private Pilot, while many were instrument rated and some were flight instructors. Total flight times ranged from 50 to 13,000 hours.

Procedures/Tasks

Pilots completed an experience questionnaire, were briefed about the experimental display, and instructed that they would recover from unknown attitudes. They were told to recover to a zero-pitch, zero-bank attitude, regardless of altitude or airspeed, as the EP would configure the aircraft such that performance was usually within the operating envelope (primary interest was in participant ability to interpret the display and determine when a

level attitude had been restored). They then entered the simulator, where they were further familiarized with the display and the simulator. A visor was used so that direct vision of the display would be obscured when in the head-down preparatory position for the recovery.

Each pilot then took off from Albuquerque, climbed into IFR conditions, and performed 8 warm-up (baseline) recovery maneuvers using the basic EADI on the PFD. Recall that the practical test standards for the Private Pilot Certificate require performance of recoveries from unknown attitudes solely by reference to instrumentation, so that all participants had received training in this task prior to participation in the experiment (recency of last recoveries was assessed by post-test questionnaire, and most reported it at their check ride or biennial flight review). Each trial began with the participant in the head-down position and hands off the controls. The EP (same for all participants) then placed the simulator into the required attitude and heading for that trial, using predetermined airspeed, altitude, and heading criteria that had been rehearsed, and told the participant to recover. Upon completing these trials, the participant flew the simulator back to the airport for a full-stop landing. At this time, the display format was changed and the procedure was repeated.

Experimental trials consisted of 16 recovery maneuvers (defined by combinations of the sampling variables described earlier) using the assigned PFD format. Two groups repeated the EADI in the experimental trials, one as a control and the other with guidance cues added. Two counterbalanced orders of the combinations of the sampling variables were used within the groups to attempt to distribute any possible learning more evenly across conditions. Pilot IRTs and TRTs were recorded for each trial. A recovery was considered complete when the aircraft reached ± 2.5 degrees of pitch and ± 5.0 degrees of bank and the pilot was able to maintain those values for 3 seconds. The supplemental trials described earlier in the Methods section were added to the end of the session. Participants completed a posttest set of questionnaires regarding their subjective assessment of the displays, went through a posttest interview, and provided both solicited and unsolicited responses/opinions.

RESULTS

Group Parity in Baseline

The distributions of hours of experience, licensing/rating categories and age were similar enough between groups that any differences found between group performances were unlikely to be a result of those variables. Analysis of recovery times for the baseline trials showed that although the groups initially differed in their performances, they

were performing comparably (no significant differences) by the last two trials, suggesting that all groups had attained a similar level of performance prior to entering the experimental trials. Means by groups and trials are depicted in Figure 5.

Comparative Analyses Approach

It is recommended that both a test for a difference and a test for equivalence be performed (Rogers, et al., 1993). The difference-test results will be presented first, followed by equivalence-test results. The ideal combination of outcomes would be that (a) performance distributions are not statistically different and (b) that they are statistically “equivalent.”

Performance Variables – Difference Tests

Recovery times. Multivariate Analysis of Variance indicated there were no significant differences between the display configurations for either of the response-time variables. Pitch-roll TRTs averaged around 10 seconds, whereas roll-only recoveries averaged about 8.5 seconds. Pitch-only recoveries averaged approximately 8.6 to 9.0 seconds. Figure 6 depicts mean TRTs by maneuver and display format. Univariate analyses were conducted to determine if type of maneuver was associated with any significant differences between group performances. Again, no significant differences were found between displays and type of maneuver for either of the response-time measures.

Control reversals. There were only three clearly identifiable primary control reversals in the nearly 800 trials. There were no secondary reversals (initial response in correct direction; subsequent control movement opposite to input required). Recovery times for the three reversals were not notably different from those of other trials. Thus, reversals did not appear to be a factor.

Supplemental trials. Analyses were conducted for performance variables on each of the three supplemental trials, and no differences in performance were found between the display configurations.

Equivalence Tests

The practice of *equivalence testing* has been used widely in the field of medicine to assess the effectiveness of new drugs or procedures as compared with existing medications or practices, particularly when a “generic” version of a medication is being introduced as a potential alternate to the established name brand. The central issue is determining if the proposed treatment or intervention is at least as good as (“equivalent to”) the existing one. Both Rogers et al. (1993) and Seaman and Serlin (1998) discuss this problem and several possible statistical approaches. The conclusion drawn is that the use of

confidence intervals requires fewer arbitrary decisions than would the use of either of two other approaches. In addition, a precedent for using this approach for performance assessment in aviation systems was set by Reising, et al. (1998) in evaluating pilot performance using pathway-display guidance for curved approaches. Also, this type of evaluation is presently being used to compare the effectiveness of flight-training devices and PC-based aviation training devices (Taylor, et al., 2004) for training and/or evaluating instrument skills (Taylor, et al., 2005; Taylor, Talleur, Rantanen, Emanuel, & Beringer, manuscript in preparation).

The procedure used by Rogers et al., was followed here. One first constructs the “equivalence interval,” a reference mean bounded by positive and negative limits about that mean representing “practical” limits within which variations in performance are not considered to be of practical significance. These can be based on baseline data or practical experience with the specific task. This can be followed by either two simultaneous one-sided hypothesis tests or the construction of confidence intervals. Using the hypothesis tests, the aim of the first test is to attempt to reject H_0 : that the difference between the two means is less than or equal to the smaller delta (lower criterion) and the second to reject an H_0 : that the difference is greater than or equal to the larger delta (upper criterion). If both null hypotheses are rejected, the comparison distribution falls within the upper and lower criterion limits.

Alternately, one can construct confidence limits for each of the comparison conditions. If the 90% confidence interval for a condition falls completely within the criterion interval, then equivalence is assumed. Additionally, if the reference mean is contained within the 95% confidence interval, then we fail to reject the null hypothesis for the traditional difference test. Thus, the target for the equivalence test is to have the 90% confidence interval within the equivalence boundaries and the reference mean within the 95% confidence interval.

Following the defined procedures, means and standard deviations for total recovery time (TRT) were generated for each participant (1) across all baseline trials and (2) across all of the experimental trials. The baseline trials were used as the basis for determining a mean reference point and for generating criterion bounds to be used in comparing confidence intervals generated for each of the display groups (the “experimental” conditions) with the reference or standard (baseline: EADI). A practical effect size was selected based upon the standard deviation of the baseline data ($sd = 1.98$; criterion chosen as 0.2 of reference mean, 1.994). Table 1 shows the means and other associated statistics used for this comparison.

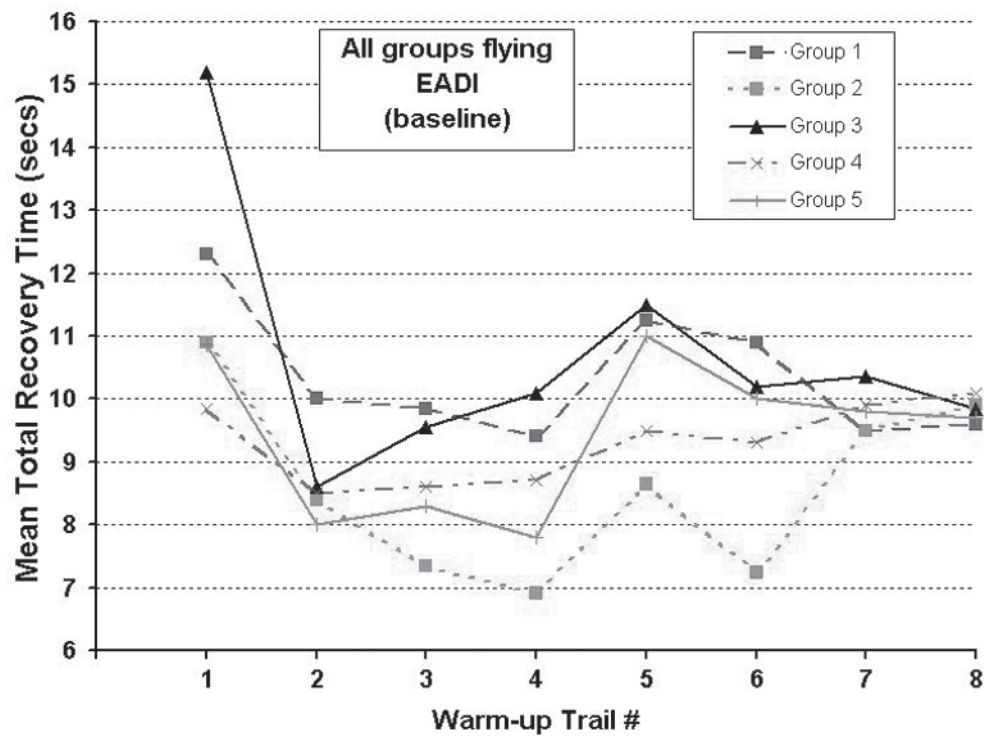


Figure 5. Mean total recovery times for baseline trials by group and serial trial number.

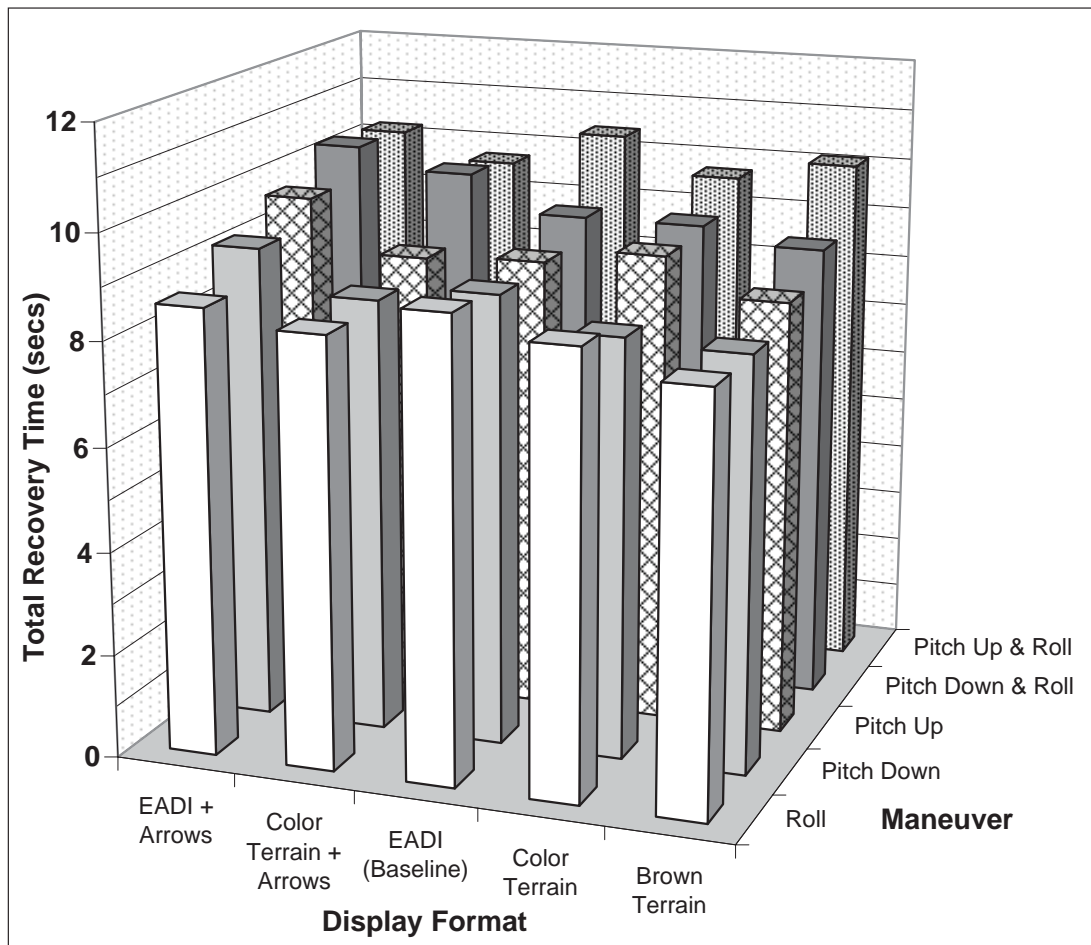


Figure 6. Mean TRT by display type and maneuver.

The philosophy for “equivalent safety” regarding this response-time measure would be that recovery times should be equal to or less than those obtained with the standard. As such, the criterion that the entire confidence interval should fall within the criterion limits for “equivalence” can be modified to the practical consideration that only the upper limit of the confidence interval is important and needs to be within the criterion limits. Correspondingly, only the z scores for the upper tails are of interest in determining that the results are “no worse than” those of the reference condition.

Examination of the p values in Table 1 associated with the upper tails of the distributions shows that all of these were within the criterion limits (we reject $H_0: m_1 - m_2 > \text{or} = \Delta_1$ and accept $H_a: \Delta_1 < m_1 - m_2 < \Delta_2$). In fact, reference to Figure 7 shows that the 90% confidence intervals for AA and AO fall completely within the criterion range, and we can accept those as functionally equivalent to our baseline.

That AO is evaluated as equivalent is particularly important in that it is identical to the baseline condition (EADI Only), suggesting that any gain in performance from baseline to experimental trials was not sufficient to be detectable as a difference, either by traditional difference testing (note that

Table 1. Reference and comparison values for the equivalence evaluation. (For Table 1 & Figure 7; AA=EADI w/arrows, AL = all features (arrows, full-color terrain), AO = EADI only, AT = EADI + terrain, BT = EADI w/brown terrain.)

	Mean	+0.2	-0.2				
Baseline (EADI)	9.87	11.96	7.98				
		90% upper	90% lower	95% upper	95% lower	upper z	lower z
AO (EADI only; same as baseline)	9.31	10.50	8.11	10.84	8.03	3.48 [#]	-1.94 ⁺
AA (EADI plus arrows)	9.67	10.84	8.49	11.17	8.21	3.04 ^x	-2.47 [*]
AT (full-color terrain)	9.01	10.22	7.81	10.55	7.68	3.87 [#]	-1.53
AL (full-color terrain plus arrows)	9.1	10.36	7.84	10.70	7.70	3.59 ^o	-1.57
BT (brown terrain)	8.78	9.91	7.66	10.23	7.54	4.47 [#]	-1.30

⁺ $p < .05$, ^{*} $p < .01$, ^x $p < .005$, ^o $p < .0005$, [#] $p < .0001$

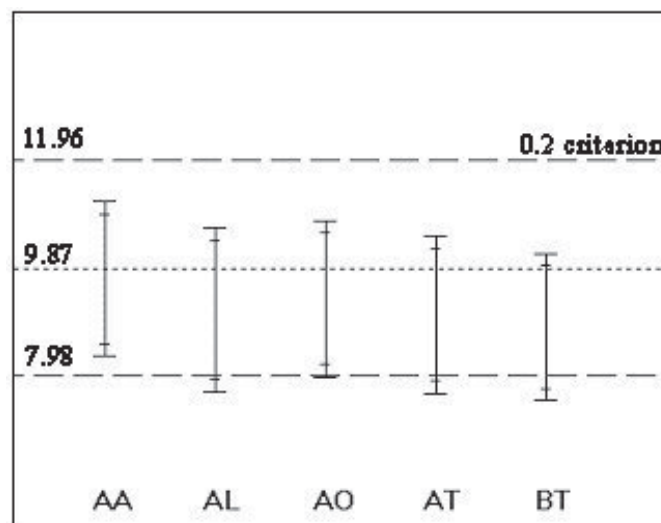


Figure 7. Experimental-condition 90% and 95% confidence intervals relative to baseline mean and upper and lower criteria graphed as seconds.

the reference mean is within the 95% confidence interval for each group) or by equivalence testing. Interestingly, AL, AT and BT, each having a terrain-background component, are well within the upper limit but are all just out of the lower limit of the criterion range (meaning a tendency towards shorter recovery times). Thus, they would be judged as no worse than the baseline and potentially (not significantly) better.

Questionnaires and Posttest Interviews

Pilots indicated that they were focusing their attention on the relatively prominent zero-pitch line and did not regard the terrain depictions as significant contributors to their recovery task. Neither preference for the 40-degree field of view nor for guidance arrows was paired with performance differences. Participants also expressed a relatively uniform preference for the terrain-depicting displays.

SUMMARY AND CONCLUSIONS

It appears that a zero-pitch line of the contrasting components and of the thickness and extent specified allows pilots to discern the zero-pitch reference from other display features and to perform recoveries from unknown attitudes regardless of the terrain format used. It also appears that the directional-guidance arrows did not produce the effect found in a previous experiment (Gershzhohn, 2001) despite a positive pilot response, and the frequency of reversals was too low to draw any conclusion from them. Although the mean total recovery times tended to be shorter for those display conditions that depicted background terrain, the differences were neither statistically nor practically significant.

Given the previous findings (indicating enhanced terrain awareness attributable to terrain depictions), combined with the lack of detrimental effects found in this study relative to recoveries from unknown attitudes, there would appear to be few significant obstacles to the implementation of this type of PFD for general aviation use. Caveats to be observed, however, would be that (1) similarly constructed terrain depictions are used, (2) the zero-pitch line is clearly differentiable from the terrain and sky depictions regardless of the type of background and (3) that the direction of off-display pitch-line locations are clearly indicated.

Further, it appears that the combination of difference testing and “equivalence” testing, using a practical difference based upon baseline data for the chosen application and task-performance variables, can provide a reasonable means by which to determine practically equivalent levels of pilot performance and, thus, equivalent levels of safety. Using this approach in the field and/or in a very applied environment will, however, always face the limitations imposed by restricted sample sizes, high performance variability, and the challenge of selecting meaningful practical effect sizes.

REFERENCES

- Alter, K.W., Barrows, A.K., Jennings, C.W., and Powell, J.D. (2000). 3-D perspective primary flight displays for aircraft. *Proceedings of the 44th Annual Meeting of the Human Factors and Ergonomics Society*, 3-29 – 3-32.
- Arthur, J.J., III, Prinzel, L.J., III, Kramer, L.J., Parrish, R.V., and Bailey, R.E. (2004). Flight simulator evaluation of synthetic vision display concepts to prevent controlled flight into terrain (CFIT). Springfield, VA: NTIS, NASA/TP-2004-213008.
- Beringer, D.B. (2000). Development of highway-in-the-sky displays for flight-path guidance: History, performance results, guidelines. *Proceedings of the 44th Annual Meeting of the Human Factors and Ergonomics Society*, 3-21 – 3-24.
- Gershzhohn, G., (2001). Unusual attitude recovery using the roll arrow. Society of Automotive Engineers, Inc., SAE Technical Paper, Document Number 2001-01-3009.
- Reising, J.M., Liggett, K.K., Kustra, T.W., and Snow, M. P. (1998). Evaluation of pathway symbology used to land from curved approaches with varying visibility conditions. *Proceedings of the 42nd Annual Meeting of the Human Factors and Ergonomics Society*, 1-5.
- Rogers, J.L., Howard, K.I., and Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, 113(3), 553-565.
- Seaman, M.A., and Serlin, R.C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods*, 3(4), 403-411.
- Taylor, H.L., Talleur, D.A., Rantanen, E.M. and Emanuel, T.W. (2004). The effectiveness of a personal computer aviation training device, a flight training device, and an airplane in conducting instrument proficiency checks. Savoy, IL: University of Illinois Institute of Aviation, Technical Report AHFD-04-12/FAA-04-5.
- Taylor, H.L., Talleur, D.A., Emanuel, T.W., and Rantanen, E. (2005). Effectiveness of flight training devices used for instrument training. Savoy, IL: University of Illinois Institute of Aviation, Final Technical Report AHFD-05-9/FAA-05-4.

